EXHIBIT D

ENGINEERING STATEMENT OF
DUTREIL, LUNDIN & RACKLEY, INC.

ENGINEERING STATEMENT
IN SUPPORT OF REPLY COMMENTS
IN RE THE REVISION OF THE CABLE
AND SATELLITE COMPULSORY LICENSES
COPYRIGHT OFFICE DOCKET No. 97-1

This Engineering Statement was prepared at the request of the Network Affiliated Stations Alliance ("NASA") for that Association's Reply Comments in Copyright Office Docket No. 97-1, In Re the Revision of the Cable and Satellite Compulsory Licenses directed toward the Satellite Home Viewer Copyright Act ("SHVA"). The SHVA includes a provision for the determination of eligibility for satellite network service to an "unserved" household based on an actual field strength measurement to determine the presence or absence of a Grade B intensity signal of the subject network station.

Grade B Definition

According to the SHVA, an "unserved" household is one that

> "…cannot receive, through the use of a conventional outdoor rooftop receiving antenna, an over-the-air signal of Grade B intensity (as defined by the Federal Communications Commission) of a primary network station affiliated with that network…"

The FCC defines the Grade B contour in Section 73.683(a) of its Rules, wherein it states:

"The required field strength, F(50,50), in decibels above one microvolt per meter (dBu) for the Grade A and Grade B contours is as follows:

	Grade B (dBu)
Channels 2-6	47
Channels 7-13	56
Channels 14-69	64

A dBu in this context is a measure of the electric field intensity of the subject signal. Section 73.686 of the FCC Rules outlines a basic procedure for measurement of TV field intensity and procedures for analysis of the measurement data. Use of this procedure results in an objective signal strength measurement that is both reliable and repeatable.

<u>Picture Quality</u>

A correlation exists between signal strength and picture quality. Consider that there are three primary components of TV transmission circuit:

1. <u>The transmission system</u> - The TV signal is transmitted through the air by the broadcaster;
2. <u>The transmission path</u> - The TV signal travels through various paths, through the atmosphere, perhaps diffracting and/or reflecting off terrain, or encountering trees, ultimately to the home;
3. <u>The receiving system</u> - The TV signal is received for viewing through a receiver.

The broadcaster controls the first component; the laws of nature control the second component; and the viewer controls the third component. If a measurement is made at a household indicating that a signal intensity of Grade B or better is present, then it indicates that components 1 and 2 of the circuit are "working" and that in most cases, with the proper receiving equipment, an acceptable picture can be obtained.

Let us examine how component 3 in the circuit affects picture quality. In the analog television domain, picture quality is a direct function of the signal-to-noise (S/N) ratio. The higher the signal level is above the noise, the better the picture quality.

Numerical estimates have been made as to the minimum S/N ratios necessary to obtain certain levels of perceived picture quality. In the late 1950's an organization was formed under the auspices of the Federal Communications Commission called the Television Allocations Study Organization (TASO). Among other tasks the TASO conducted extensive television picture quality assessment experiments to develop a correlation between the S/N ratio and perceived picture quality.[*] The Federal Communications Commission in 1960 further analyzed these data.[†] This analysis showed the following results:

---

[*] Television Allocations Study Organization (TASO), Engineering Aspects of Television Allocations, Report to the Federal Communications Commission, March 16, 1959.

[†] Fine, Harry, A Further Analysis of TASO Panel 6 Data on Signal to Interference Ratios and Their Application to Description of Television Service, FCC/OCE Report T.R.R. Report No. 5.1.2., April 1, 1960.

S/N (dB)	Picture Quality	TASO Grade
> 41	Excellent	1
33-41	Fine	2
28-33	Passable	3
23-28	Marginal	4
17-23	Inferior	5
< 17	Unusable	6

The FCC Grade B contour was defined based on a S/N ratio of 30 dB[‡], which is equivalent to a passable (or acceptable) picture quality or an approximate TASO Grade 3.[§]

The following explains how the standard may be applied. Now consider that the received S/N is a function of the following:

$$S/N = E + K + G - F - N_i, \text{ where}$$

E = intensity of the television signal in the vicinity of the receive antenna, expressed in dBu
F = system noise figure in dB (A combination of noise factors including transmission line loss, television receiver noise figure, etc.
N_i = inherent thermal noise voltage generated at the input of an ideal receiver. For standard TV sets this is approximately 7 dB.
K = dipole factor for frequency of interest in dB.
G = gain of the receiving antenna referenced to a half-wave dipole in dB.[**]

When the FCC developed the television service, it used certain planning factors to estimate television coverage.

‡ A S/N ratio of 30 dB had been established in the 1940s as the threshold for acceptable service. (See, for example, George H. Brown, "Field Test of Ultra-High-Frequency Television in the Washington Area", RCA Review, Vol. IX, No. 4, pp. 565-584; December, 1948.)
§ The TASO picture quality assessments involved ten groups of observers, each consisting of nominally 20 persons (See Engineering Aspects of Television Allocations)

As an example, let us examine the picture quality that would obtain under the assumption of the FCC's planning factors for a Channel 38 facility:

Typical outdoor receiving antenna	Gain (G) = 13 dB at Channel 38
Transmission line	Loss = 5 dB
Typical TV set	Noise Figure = 15 dB

The system noise figure would be equal to 5 dB plus 15 dB, or 20 dB in this case. The calculated S/N using these figures would be as follows:

$$S/N = 64 + (-16.1) + 13 - 20 - 7$$
$$S/N = 33.9 \text{ dB.}$$

Under these assumptions, the S/N ratio would be 33.9 dB and the viewer would obtain a picture quality level in the "Fine" range under the TASO grading standard.

The Federal Communications Commission recognized the potential for the TV signal to vary with time so a factor of 4 dB was added to the UHF planning factors to mitigate the effects of time variability. In other words, in this example, if the signal dropped an additional 4 dB as a result of propagational effects, the viewer would receive a S/N ratio of 29.9 dB or a TASO Grade 3 "Passable" picture quality. Calculations of the S/N ratio obtained for low VHF (Channel 2-6), high VHF (Channels 7-13) and UHF stations under the FCC's assumptions for Grade B coverage are summarized below:

** Felker, Alex D., UHF Reception and Television Preamplifiers, FCC Office of Plans and Policy, April 1981.

Channels 2-6	Measured Signal Intensity = 47 dBu Antenna Gain (G) = 6 dB System Noise Figure (F) = 13 dB Dipole Factor (K) = +3	S/N = 47 + 3 + 6 – 13 – 7 S/N = 36 dB TASO Grade = 2 (Fine)
Channels 7-13	Measured Signal Intensity = 56 dBu Antenna Gain (G) = 6 dB System Noise Figure (F) = 14 dB Dipole Factor (K) = -6	S/N = 56 + (-6) + 6 – 14 – 7 S/N = 35 dB TASO Grade = 2 (Fine)
Channels 7-13	Measured Signal Intensity = 64 dBu Antenna Gain (G) = 13 dB System Noise Figure (F) = 20 dB Dipole Factor (K) = -16	S/N = 64 + (-16) + 13 – 20 – 7 S/N = 34 dB TASO Grade = 2 (Fine)

Thus, as indicated above, if a median signal of Grade B intensity is measured, it will translate into a TASO Grade 2 "Fine" picture for low VHF, high VHF and UHF stations. There are fade margins of 6 dB, 5 dB and 4 dB, for low VHF, high VHF and UHF stations, respectively to a S/N level of 30 dB, which is equivalent to a passable quality picture under the TASO Grades.

Noise

The question has been raised whether the presence of man-made and other environmental noise might degrade picture quality. In some circumstances it can. However, external environmental noise only affects VHF television stations, and it is has the greatest effect on low band VHF stations. External environmental noise does not adversely affect the picture quality of UHF stations.

Man-made noise is less prevalent in rural areas and is more likely to be a factor for VHF stations in populated urban areas. However, the signal strength of the local stations in such areas is likely to be far in excess of the Grade B level and thus sufficient to overcome the adverse effects of the noise on picture quality.

In fact, while there may be instances where household noise is present that would affect a VHF station, often times there are means to mitigate such noise. For example, the use of a higher gain antenna or a pre-amplifier might be employed to minimize the effects.

Ghosting or Multipath

What is known as television picture "ghosting", is the result of multipath, or multiple signals arriving at the receiver at different times. This effect can occur on all television channels. In many cases, the "ghosting" effect can be reduced or eliminated through the use of an improved antenna with a high front-to-back ratio. That is, because the reflected signals often arrive at different angles off the receive antenna than from the main beam of the antenna, an antenna with a high front-to-back ratio can be used to attenuate, or reduce, the level of reflected signals while preserving the desired signal. Commonly available log-periodic antennas are known for their high front-to-back ratio. A UHF "bow-tie" antenna with a wire-grid reflector is another good example of a UHF antenna with a high front-to-back ratio.

Subjective Picture Quality Assessment

A recommendation has been made by PrimeTime 24 that the viewer himself or herself grade the received picture on his or her own receiving system and determine if it is acceptable. Such a procedure would be inconsistent with good engineering practice. First, the household might have a sub-standard receiving system. Second, the receiving system may not be functioning properly. Third, the antenna may not be oriented for optimum reception of the subject station. Fourth, the viewer may be biased.

If a subjective picture quality assessment procedure were to be applied, then industry accepted practices should be employed in order to eliminate discrepancies in equipment, viewing conditions, and personal bias. Such a methodology has been prescribed by the International Telecommunications Union (ITU) in the form of Recommendation ITU-R BT.500-7 (1995)[††]. This methodology, a copy of which is included herein as Appendix I, outlines the proper procedure for the subjective assessment of television picture quality.

According to the ITU Recommendation, the proper assessment of picture quality involves such things as the careful arrangement of viewing conditions, and the selection of multiple observers with particular characteristics. In order to obtain comparable results the viewing conditions must be carefully arranged. For example, it is necessary to control the brightness and contrast of the television picture. Also, the brightness of ambient light behind the television monitor must be kept below a certain level. Other room illumination should be kept to a low level. And, the viewers' observation angle with respect to the television screen should not

†† ITU Recommendation ITU-R 500.3-7, "Method for the Subjective Assessment of the Quality of Television Pictures," Recommendations and Reports of the ITU, 1995.

exceed 30° relative to the normal. There are also specified viewing distances which must be maintained for the type of assessment method employed. The ITU recommendation requires at least 15 non-expert observers. The observers should be screened for normal visual acuity and for normal color vision.

The double stimulus continuous quality method is prescribed for this type of assessment whereby the observers are asked to compare a series of randomly presented pairs of pictures, one of which would be the television picture under examination. The observers would grade the two pictures based on a five-point quality scale. This procedure is detailed in Section 5 of the ITU Recommendation.

Conclusion

The Grade B signal strength standard is based on a picture quality standard. There is a correlation between signal strength and picture quality. The stronger the signal the better the potential quality of the picture.

Interference and multipath reflections can degrade picture quality. But through the use of improved antennas, many of these problems can be reduced or eliminated.

A picture quality assessment procedure requires multiple viewers and careful control of the viewing environment to obtain valid results.

Louis Robert du Treil

Louis Robert du Treil, Jr., P.E.

du Treil, Lundin & Rackley, Inc.
240 N. Washington Blvd., Suite 700
Sarasota, FL 34236

June 17, 1997

Appendix I

ENGINEERING STATEMENT
IN SUPPORT OF REPLY COMMENTS
IN RE THE REVISION OF THE CABLE
AND SATELLITE COMPULSORY LICENSES
COPYRIGHT OFFICE DOCKET No. 97-1

{attachment}

SECTION 11E: QUALITY ASSESSMENTS

RECOMMENDATION ITU-R BT.500-7

METHODOLOGY FOR THE SUBJECTIVE ASSESSMENT OF THE QUALITY OF TELEVISION PICTURES

(Question ITU-R 211/11)

(1974-1978-1982-1986-1990-1992-1994-1995)

The ITU Radiocommunication Assembly,

considering

a) that a large amount of information has been collected about the methods used in various laboratories for the assessment of picture quality;

b) that examination of these methods shows that there exists a considerable measure of agreement between the different laboratories about a number of aspects of the tests;

c) that the adoption of standardized methods is of importance in the exchange of information between various laboratories;

d) that routine or operational assessments of picture quality and/or impairments using a five-grade quality and impairment scale made during routine or special operations by certain supervisory engineers, can also make some use of certain aspects of the methods recommended for laboratory assessments;

e) that the introduction of new kinds of television signal processing such as digital coding and bit-rate reduction, new kinds of television signals using time-multiplexed components and, possibly, new services such as enhanced television and HDTV may require changes in the methods of making subjective assessments;

f) that the introduction of such processing, signals and services, will increase the likelihood that the performance of each section of the signal chain will be conditioned by processes carried out in previous parts of the chain,

recommends

1 that the general methods of test, the grading scales and the viewing conditions for the assessment of picture quality, described in the following Annexes should be used for laboratory experiments and whenever possible for operational assessments;

2 that, in the near future and notwithstanding the existence of alternative methods and the development of new methods, those described in § 4 and 5 of Annex 1 to this Recommendation should be used when possible;

3 that, in view of the importance of establishing the basis of subjective assessments, the fullest descriptions possible of test configurations, test materials, observers, and methods should be provided in all test reports;

4 that, in order to facilitate the exchange of information between different laboratories, the collected data should be processed in accordance with the statistical techniques detailed in Annex 2 to this Recommendation.

NOTE 1 – Information on subjective assessment methods for establishing the performance of television systems is given in Annex 1.

NOTE 2 – Description of statistical techniques for the processing of the data collected during the subjective tests is given in Annex 2.

ANNEX 1

Description of assessment methods

1 Introduction

Subjective assessment methods are used to establish the performance of television systems using measurements that more directly anticipate the reactions of those who might view the systems tested. In this regard, it is understood that it may not be possible to fully characterize system performance by objective means; consequently, it is necessary to supplement objective measurements with subjective measurements.

In general, there are two classes of subjective assessments. First, there are assessments that establish the performance of systems under optimum conditions. These typically are called quality assessments. Second, there are assessments that establish the ability of systems to retain quality under non-optimum conditions that relate to transmission or emission. These typically are called impairment assessments.

To conduct appropriate subjective assessments, it is first necessary to select from the different options available those that best suit the objectives and circumstances of the assessment problem at hand. To help in this task, after the general features reported in § 2, some information is given in § 3 on the assessment problems addressed by each method. Then, the two main recommended methods are detailed in § 4 and 5. Finally, general information on alternative methods under study is reported in § 6.

The purpose of this Annex is limited to the detailed description of the assessment methods. The choice of the most appropriate method is nevertheless dependent on the service objectives the system under test aims at. The complete evaluation procedures of specific applications are therefore reported in other ITU-R Recommendations.

2 Common features

2.1 General viewing conditions

The assessors' viewing conditions should be arranged as follows:

a)	Ratio of luminance of inactive screen to peak luminance:	≤ 0.02
b)	Ratio of the luminance of the screen, when displaying only black level in a completely dark room, to that corresponding to peak white:	≈ 0.01
c)	Display brightness and contrast:	set up via PLUGE (see Recommendations ITU-R BT.814 and ITU-R BT.815)
d)	Maximum observation angle relative to the normal. (This number applies to CRT displays, whereas the appropriate numbers for other displays are under study.):	30°
e)	Ratio of luminance of background behind picture monitor to peak luminance of picture:	≈ 0.15
f)	Chromaticity of background:	D_{65}
g)	Other room illumination:	low

The viewing distance, (Note 1) the maximum observation angle and the peak luminance of the screen are dependent on the application which has to be quantified. Therefore, the appropriate values are reported in the Recommendations addressing the application under test.

NOTE 1 – The application dependent design viewing distance is recommended but in some cases, such as home viewing, another concept called preferred viewing distance could be used.

It has been noted that, when left to their own devices, viewers may elect for viewing distances greater than those used in subjective assessments. The relationship between preferred viewing distances and those used in assessments needs further study.

2.2 Source signals

The source signal provides the reference picture directly, and the input for the system under test. It should be of optimum quality for the television standard used. The absence of defects in the reference part of the presentation pair is crucial to obtaining stable results.

Digitally stored pictures and sequences are the most reproducible source signals, and these are therefore the preferred type. They can be exchanged between laboratories, to make system comparisons more meaningful. Video or computer tapes are possible formats.

In the short term, 35 mm slide-scanners provide a preferred source for still pictures. The resolution available is adequate for evaluation of conventional television. The colorimetry and other characteristics of film may give a different subjective appearance to studio camera pictures. If this affects the results, direct studio sources should be used, although this is often much less convenient. As a general rule, slide-scanners should be adjusted picture by picture for best possible subjective picture quality, since this would be the situation in practice.

Assessments of downstream processing capacity are often made with colour-matte. In studio operations, colour-matte is very sensitive to studio lighting. Assessments should therefore preferably use a special colour-matte slide pair, which will consistently give high-quality results. Movement can be introduced into the foreground slide if needed.

It will be frequently required to take account of the manner in which the performance of the system under test may be influenced by the effect of any processing that may have been carried out at an earlier stage in the history of the signal. It is therefore desirable that whenever testing is carried out on sections of the chain that may introduce processing distortions, albeit non-visible, the resulting signal should be transparently recorded, and then made available for subsequent tests downstream, when it is desired to check how impairments due to cascaded processing may accumulate along the chain. Such recordings should be kept in the library of test material, for future use as necessary, and include with them a detailed statement of the history of the recorded signal.

2.3 Selection of test materials

A number of approaches have been taken in establishing the kinds of test material required in television assessments. In practice, however, particular kinds of test materials should be used to address particular assessment problems. A survey of typical assessment problems and of test materials used to address these problems is given in Table 1.

TABLE 1

Selection of test material*

Assessment problem	Material used
Overall performance with average material	General, "critical but not unduly so"
Capacity, critical applications (e.g. contribution, post-processing, etc.)	Range, including very critical material for the application tested
Performance of "adaptive" systems	Material very critical for "adaptive" scheme used
Identify weaknesses and possible improvements	Critical, attribute-specific material
Identify factors on which systems are seen to vary	Wide range of very rich material
Conversion among different standards	Critical for differences (e.g. field rate)

* It is understood that all test materials could conceivably be part of television programme content. For further guidance on the selection of test materials, see Appendices 1 and 2 to Annex 1.

Some parameters may give rise to a similar order of impairments for most pictures or sequences. In such cases, results obtained with a very small number of pictures or sequences (e.g. two) may still provide a meaningful evaluation.

However, new systems frequently have an impact which depends heavily on the scene or sequence content. In such cases, there will be, for the totality of programme hours, a statistical distribution of impairment probability and picture or sequence content. Without knowing the form of this distribution, which is usually the case, the selection of test material and the interpretation of results must be done very carefully.

In general, it is essential to include critical material, because it is possible to take this into account when interpreting results, but it is not possible to extrapolate from non-critical material. In cases where scene or sequence content affects results, the material should be chosen to be "critical but not unduly so" for the system under test. The phrase "not unduly so" implies that the pictures could still conceivably form part of normal programme hours. At least four items should, in such cases, be used: for example, half of which are definitely critical, and half of which are moderately critical.

A number of organizations have developed test still pictures and sequences. It is hoped to organize these in the framework of the ITU-R in the future. Specific picture material is proposed in the Recommendations addressing the evaluation of the applications.

Further ideas on the selection of test materials are given in Appendices 1 and 2.

2.4 Range of conditions and anchoring

Because most of the assessment methods are sensitive to variations in the range and distribution of conditions seen, judgement sessions should include the full ranges of the factors varied. However, this may be approximated with a more restricted range, by presenting also some conditions that would fall at the extremes of the scales. These may be represented as examples and identified as most extreme (direct anchoring) or distributed throughout the session and not identified as most extreme (indirect anchoring).

2.5 Observers

At least 15 observers should be used. They should be non-expert, in the sense that they are not directly concerned with television picture quality as part of their normal work, and are not experienced assessors (Note 1). Prior to a session, the observers should be screened for (corrected-to-) normal visual acuity on the Snellen or Landolt chart, and for normal colour vision using specially selected charts (Ishihara, for instance). The number of assessors needed depends upon the sensitivity and reliability of the test procedure adopted and upon the anticipated size of the effect sought.

NOTE 1 – Preliminary findings suggest that non-expert observers may yield more critical results with exposure to higher quality transmission and display technologies.

2.6 Instructions for the assessment

Assessors should be carefully introduced to the method of assessment, the types of impairment or quality factors likely to occur, the grading scale, the sequence and timing. Training sequences demonstrating the range and the type of the impairments to be assessed should be used with illustrating pictures other than those used in the test, but of comparable sensitivity. In the case of quality assessments, quality may be defined as to consist of specific perceptual attributes.

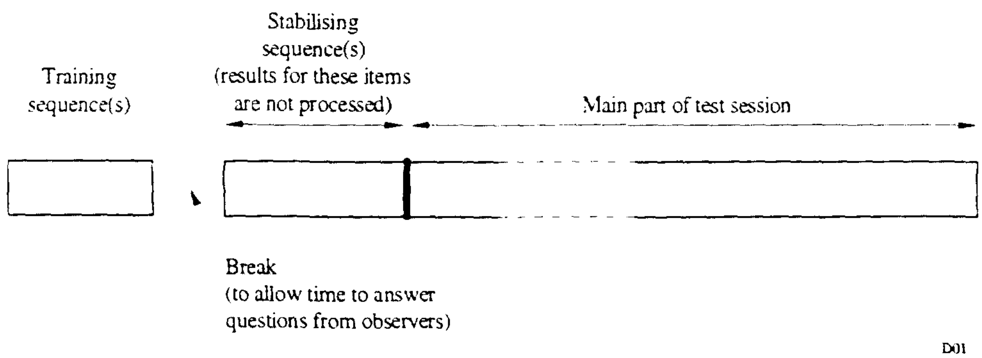
2.7 The test session

A session should last up to half an hour. At the beginning of the first session, about five "dummy presentations" should be introduced to stabilize the observers' opinion. The data issued from these presentations must not be taken into account in the results of the test. If several sessions are necessary, about three dummy presentations are only necessary at the beginning of the following session.

A random order should be used for the presentations (for example, derived from Graeco-Latin squares); but the test condition order should be arranged so that any effects on the grading of tiredness or adaptation are balanced out from session to session. Some of the presentations can be repeated from session to session to check coherence.

FIGURE 1

Presentation structure of test session



2.8 Presentation of the results

Because they vary with range, it is inappropriate to interpret judgements from most of the assessment methods in absolute terms (e.g. the quality of an image or image sequence).

For each test parameter, the mean and 5% confidence interval of the statistical distribution of the assessment grades must be given. If the assessment was of the change in impairment with a changing parameter value, curve-fitting techniques should be used. Logistic curve-fitting and logarithmic axis will allow a straight line representation, which is the preferred form of presentation. More information on data processing is given in Annex 2 to this Recommendation.

The results must be given together with the following information:

- details of the test configuration,
- details of the test materials,
- type of picture source and display monitors,
- number and type of assessors,
- reference systems used,
- the grand mean score for the experiment,
- original and adjusted mean scores and 5% confidence interval if one or more observers have been eliminated according to the procedure given below.

3 Selection of test methods

A wide variety of basic test methods have been used in television assessments. In practice, however, particular methods should be used to address particular assessment problems. A survey of typical assessment problems and of methods used to address these problems is given in Table 2.

4 The double-stimulus impairment scale method (the "EBU method")

4.1 General description

A typical assessment might call for an evaluation of either a new system, or the effect of a transmission path impairment. The initial steps for the test organizer would include the selection of sufficient test material to allow a meaningful evaluation to be made, and the establishment of which test conditions should be used. If the effect of parameter variation is of interest, it is necessary to choose a set of parameter values which cover the impairment grade range in a small number of roughly equal steps. If a new system, for which the parameter values cannot be so varied, is being evaluated, then either additional, but subjectively similar, impairments need to be added, or another method such as that in § 5 should be used.

TABLE 2

Selection of test methods

Assessment problem	Method used	Description
Measure the quality of systems relative to a reference	Double stimulus continuous quality method	Rec. ITU-R BT.500 § 5
Measure the robustness of systems (i.e. failure characteristics)	Double stimulus impairment method	Rec. ITU-R BT.500 § 4
Quantify the quality of systems (when no reference is available)	Ratio-scaling method(1) or categorical scaling, under study	Report ITU-R BT.1082
Compare the quality of alternative systems (when no reference is available)	Method of direct comparison, ratio-scaling method(1) or categorical scaling, under study	Report ITU-R BT.1082
Identify factors on which systems are perceived to differ and measure their perceptual influence	Method under study	Report ITU-R BT.1082
Establish the point at which an impairment becomes visible	Threshold estimation by forced-choice method or method of adjustment, under study	Report ITU-R BT.1082
Determine whether systems are perceived to differ	Forced-choice method, under study	Report ITU-R BT.1082

(1) Some studies suggest that this method is more stable when a full range of quality is available.

The double-stimulus (EBU) method is cyclic in that the assessor is first presented with an unimpaired reference, then with the same picture impaired. Following this, he is asked to vote on the second, keeping in mind the first. In sessions, which last up to half an hour, the assessor is presented with a series of pictures or sequences in random order and with random impairments covering all required combinations. The unimpaired picture is included in the pictures or sequences to be assessed. At the end of the series of sessions, the mean score for each test condition and test picture is calculated.

The method uses the impairment scale, for which it is usually found that the stability of the results is greater for small impairments than for large impairments. Although the method sometimes has been used with limited ranges of impairments, it is more properly used with a full range of impairments.

4.2 General arrangement

The viewing conditions, source signals, test material, the observers and the presentation of results are defined or selected in accordance with § 2.

The generalized arrangement for the test system should be as shown in Fig. 2.

The assessors view an assessment display which is supplied with a signal via a timed switch. The signal path to the timed switch can be either directly from the source signal or indirectly via the system under test. Assessors are presented with a series of test pictures or sequences. They are arranged in pairs such that the first in the pair comes direct from the source, and the second is the same picture via the system under test.

4.3 Presentation of the test material

A test session comprises a number of presentations. There are two variants to the structure of presentations, I) and II) outlined below.

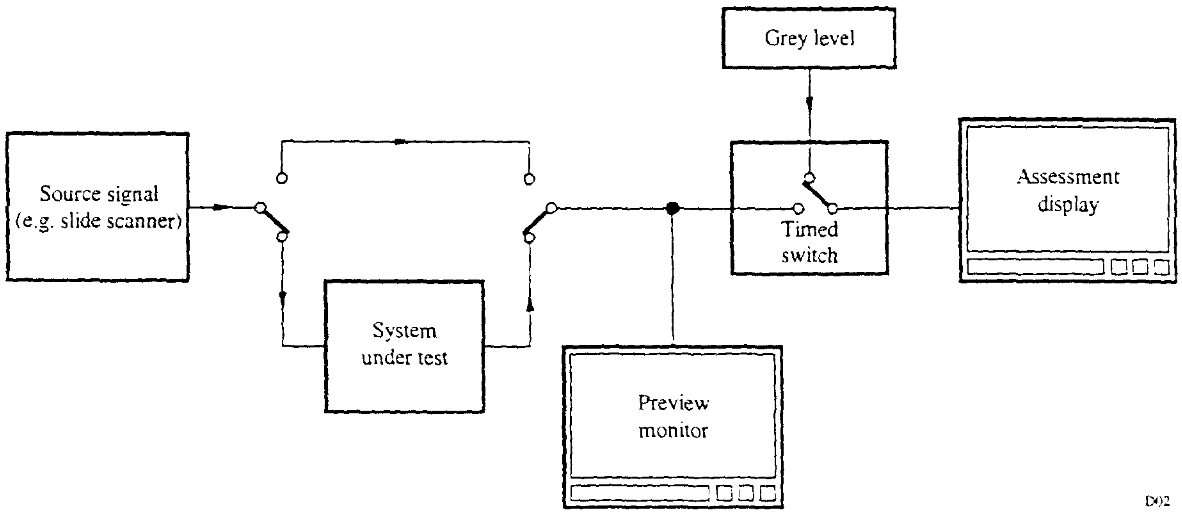
Variant I: The reference picture or sequence and the test picture or sequence are presented only once as is shown in Fig. 3a).

Variant II: The reference picture or sequence and the test picture or sequence are presented twice as is shown in Fig. 3b).

Variant II, which is more time consuming than variant I, may be applied if the discrimination of very small impairments is required or moving sequences are under test.

FIGURE 2

General arrangement for test system for double-stimulus impairment scale method



4.4 Grading scales

The five-grade impairment scale should be used:

5 imperceptible

4 perceptible, but not annoying

3 slightly annoying

2 annoying

1 very annoying

Assessors should use a form which gives the scale very clearly, and has numbered boxes or some other means to record the gradings.

4.5 The introduction to the assessments

At the beginning of each session, an explanation is given to the observers about the type of assessment, the grading scale, the sequence and timing (reference picture, grey, test picture, voting period). The range and type of the impairments to be assessed should be illustrated on pictures other than those used in the tests, but of comparable sensitivity. It must not be implied that the worst quality seen necessarily corresponds to the lowest subjective grade. Observers should be asked to base their judgement on the overall impression given by the picture, and to express these judgements in terms of the wordings used to define the subjective scale.

The observers should be asked to look at the picture for the whole of the durations of T1 and T3. Voting should be permitted only during T4.

4.6 The test session

The pictures and impairments should be presented in a pseudo-random sequence and, preferably in a different sequence for each session. In any case, the same test picture or sequences should never be presented on two successive occasions with the same or different levels of impairment.

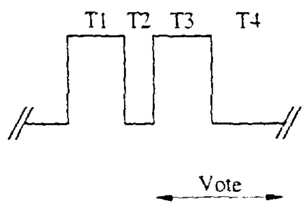
The range of impairments should be chosen so that all grades are used by the majority of observers: a grand mean score (averaged overall judgements made in the experiment) close to three should be aimed at.

A session should not last more than roughly half an hour, including the explanations and preliminaries: the test sequence could begin with a few pictures indicative of the range of impairments; judgements of these pictures would not be taken into account in the final results.

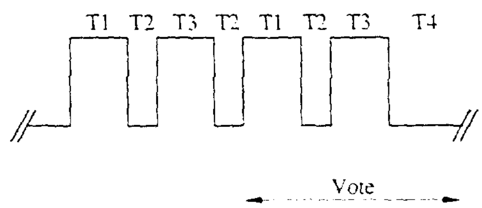
Further ideas on the selection of levels of impairments are given in Appendix 2.

FIGURE 3

Presentation structure of test material



a) Variant I



b) Variant II

Phases of presentation:

T1 =	10 s	Reference picture
T2 =	3 s	Mid grey produced by a video level of around 200 mV
T3 =	10 s	Test condition
T4 =	5-11 s	Mid grey

Experience suggests that extending the periods T1 and T3 beyond 10 s does not improve the assessors' ability to grade the pictures or sequences.

D03

5 The double-stimulus continuous quality-scale method

5.1 General description

A typical assessment might call for evaluation of a new system or of the effects of transmission paths on quality. The double-stimulus method is thought to be especially useful when it is not possible to provide test stimulus test conditions that exhibit the full range of quality.

The method is cyclic in that the assessor is asked to view a pair of pictures, each from the same source, but one via the process under examination, and the other one directly from the source. He is asked to assess the quality of both.

In sessions which last up to half an hour, the assessor is presented with a series of picture pairs (internally random) in random order, and with random impairments covering all required combinations. At the end of the sessions, the mean scores for each test condition and test picture are calculated.

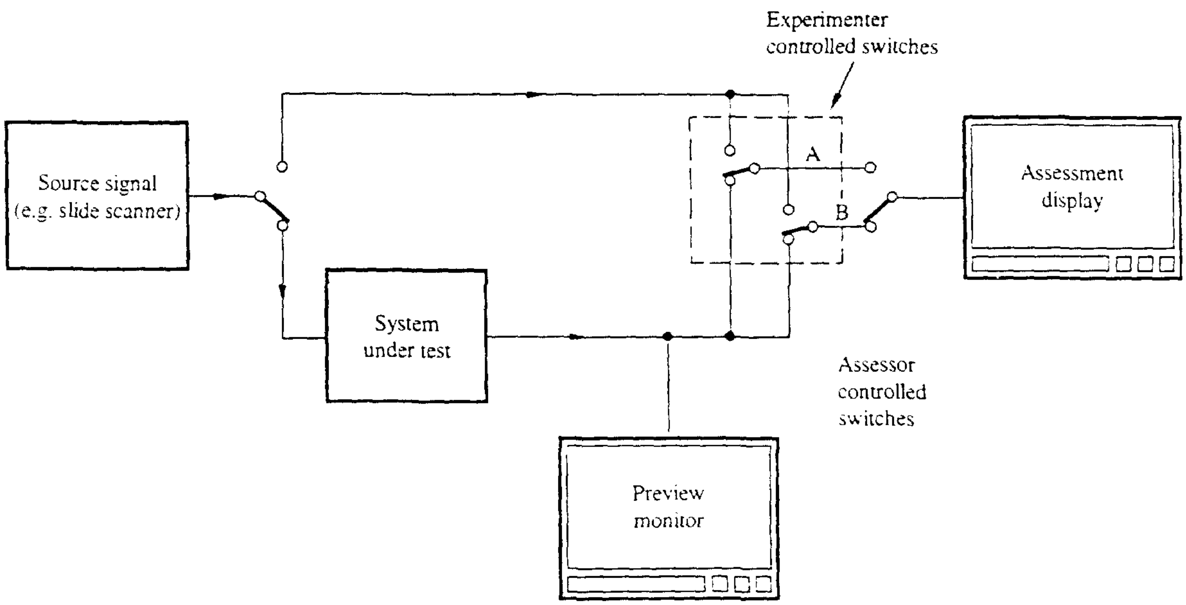
5.2 General arrangement

The viewing conditions, source signals, test material, the observers and the introduction to the assessment are defined or selected in accordance with § 2. The test session is as described in § 4.6.

The generalized arrangement for the test system should be as shown in Fig. 4 below.

FIGURE 4

General arrangement for test system for double-stimulus continuous quality-scale method



There are two variants to this method, (I) and (II), outlined below.

(I) The assessor, who is normally alone, is allowed to switch between two conditions A and B until he is satisfied that he has established his opinion of each. The A and B lines are supplied with the reference direct picture, or the picture via the system under test, but which is fed to which line is randomly varied between one test condition and the next, noted by the experimenter, but not announced.

(II) The assessors are shown consecutively the pictures from the A and B lines, to establish their opinion of each. The A and B lines are fed for each presentation as in variant (I) above. The stability of results of this variant with a limited range of quality is considered to be still under investigation.

D04

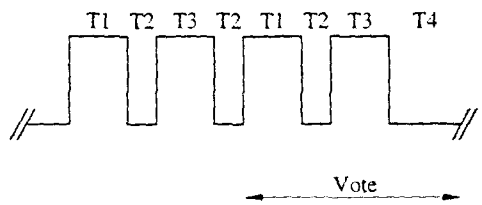
5.3 Presentation of the test material

A test session comprises a number of presentations. For variant (I) which has a single observer, for each presentation the assessor is free to switch between the A and B signals until the assessor has the mental measure of the quality associated with each signal. The assessor may typically choose to do this two or three times for periods of up to 10 s. For variant (II) which uses a number of observers simultaneously, prior to recording results, the pair of conditions is shown one or more times for an equal length of time to allow the assessor to gain the mental measure of the qualities associated with them, then the pair is shown again one or more times while the results are recorded. The number of repetitions depends on the length of the test sequences. For still pictures, a 3-4 s sequence and five repetitions (voting during the last two) may be appropriate. For moving pictures with time-varying artefacts, a 10 s sequence with two repetitions (voting during the second) may be appropriate. The structure of presentations is shown in Fig. 5.

Where practical considerations limit the duration of sequences available to less than 10 s, compositions may be made using these shorter sequences as segments, to extend the display time to 10 s. In order to minimize discontinuity at the joints, successive sequence segments may be reversed in time (sometimes called "palindromic" display). Care must be taken to ensure that test conditions displayed as reverse time segments represent causal processes, that is, they must be obtained by passing the reversed-time source signal through the system under test.

FIGURE 5

Presentation structure of test material



Phases of presentation:

T1 =	10 s	Test sequence A
T2 =	3 s	Mid grey produced by a video level of around 200 mV
T3 =	10 s	Test sequence B
T4 =	5-11 s	Mid grey

D05

5.4 Grading scale

The method requires the assessment of two versions of each test picture. One of each pair of test pictures is unimpaired while the other presentation might or might not contain an impairment. The unimpaired picture is included to serve as a reference, but the observers are not told which is the reference picture. In the series of tests, the position of the reference picture is changed in pseudo-random fashion.

The observers are simply asked to assess the overall picture quality of each presentation by inserting a mark on a vertical scale. The vertical scales are printed in pairs to accommodate the double presentation of each test picture. The scales provide a continuous rating system to avoid quantizing errors, but they are divided into five equal lengths which correspond to the normal ITU-R five-point quality scale. The associated terms categorizing the different levels are the same as those normally used; but here they are included for general guidance and are printed only on the left of the first scale in each row of ten double columns on the score sheet. Figure 6 shows a section of a typical score sheet. Any possibility of confusion between the scale divisions and the test results is avoided by printing the scales in blue and recording the results in black.

5.5 Presentation of the results

The general information about presentation of the results, mentioned in § 2.8 and Annex 2 does apply.

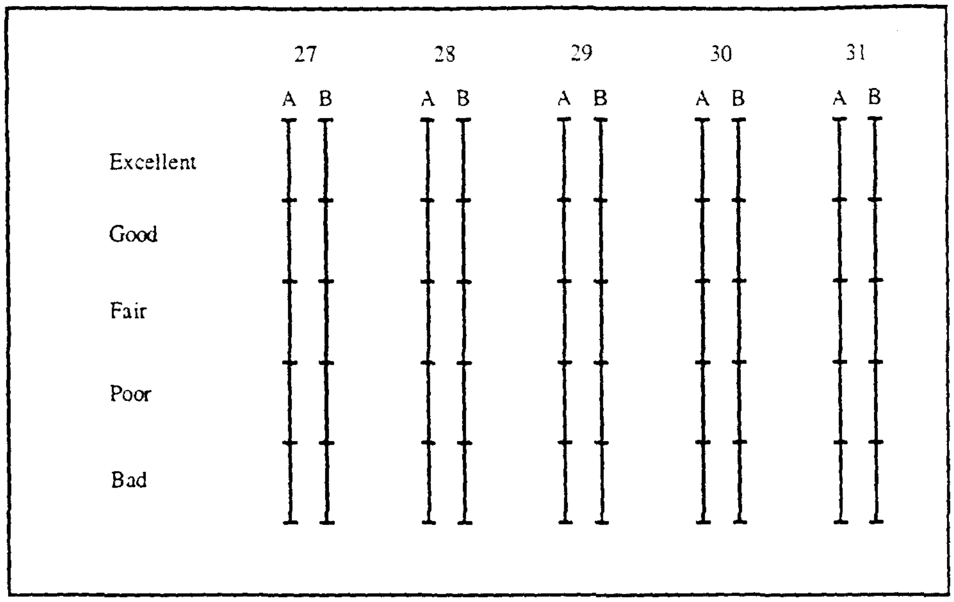
Two different approaches are possible:

– First, the results can be expressed in the form of a comparison test, i.e. to indicate directly the change in quality from the reference condition. For each test parameter, the mean and 5% confidence interval of the statistical distribution of the measured difference must be given.

– Second (the preferred presentation method), the results can be converted into the terms used to describe an equivalent quality grade. The pairs of assessments (reference and test) for each separate test condition are converted from measurements of length on the score sheet to normalized scores in the range 0 to 100. For each system under test, these scores are then averaged for the different groups of observers, different viewing distances and different test pictures, to give mean scores for reference and test conditions for each combination of the variables.

FIGURE 6

Portion of quality-rating form using continuous scales*



* In planning the arrangement of test items within a test session for the Double Stimulus Continuous Quality Scale Method it is desirable that the experimenter should include checks to give confidence that the experiment is free of systematic errors. However, the method for performing these confidence checks is under investigation.

D06

Because the mean scores for the reference conditions are always less than 1.0, a re-scaling operation on the test scores is necessary. The re-scaling is effected by subtracting residual impairment. The mean score for the reference condition is treated as the residual impairment. The results of the subtraction are expressed in impairment units (imps) but can be transformed back to mean scores if so desired.

In cases where re-scaling is not used, experimenters should note that when assessing test material which has a low quality "reference" the effective portion of the DSCQ scale range that is available for assessors to record degradations relative to the reference will be restricted. For this reason caution should be exercised in comparing or combining data for low quality reference conditions with data obtained for test sequences with relatively high quality reference conditions.

6 Alternative methods of assessment

In appropriate circumstances, the single-stimulus and stimulus-comparison methods should be used.

6.1 Single-stimulus methods

In single-stimulus methods, a single image or sequence of images is presented and the assessor provides an index of the entire presentation.

6.1.1 General arrangement

The way viewing conditions, source signals, range of conditions and anchoring, the observers, the introduction to the assessment and the presentation of the results are defined or selected is in accordance with § 2.

6.1.2 Selection of test material

For laboratory tests, the content of the test images should be selected as described in § 2.3.

Once the content is selected, test images are prepared to reflect the design options under consideration or the range(s) of one (or more) factors. When two or more factors are examined, the images can be prepared in two ways. In the first, each image represents one level of one factor only. In the other, each image represents one level of every factor examined but, across images, each level of every factor occurs with every level of all other factors. Both methods permit results to be attributed clearly to specific factors. The latter method also permits the detection of interactions among factors (i.e. non-additive effects).

6.1.3 Test session

The session consists of a series of assessment trials. These should be presented in random sequence and, preferably, in a different random sequence for each observer. When a single random sequence is used, the experimenter normally ensures that the same image is not presented twice in succession with the same kind and level of impairment.

A typical assessment trial consists of three displays: a mid-grey adaptation field, a stimulus field, and a mid-grey post-exposure field. The durations of these displays vary with viewer task, materials (e.g. still vs. moving), and the options or factors considered, but 3, 10 and 10 s respectively, are not uncommon. The viewer index, or indices, may be collected during display of either the stimulus or the post-exposure field.

6.1.4 Types of single-stimulus methods

In general, three types of single-stimulus methods have been used in television assessments.

6.1.4.1 Adjectival categorical judgement methods

In adjectival categorical judgements, observers assign an image or image sequence to one of a set of categories that, typically, are defined in semantic terms. The categories may reflect judgements of whether or not an attribute is detected (e.g. to establish the impairment threshold). Categorical scales that assess image quality and image impairment, have been used most often, and the ITU-R scales are given in Table 3 below. In operational monitoring, half grades sometimes are used. Scales that assess text legibility, reading effort, and image usefulness have been used in special cases.

TABLE 3

ITU-R quality and impairment scales

Five-grade scale			
Quality		Impairment	
5	Excellent	5	Imperceptible
4	Good	4	Perceptible, but not annoying
3	Fair	3	Slightly annoying
2	Poor	2	Annoying
1	Bad	1	Very annoying

This method yields a distribution of judgements across scale categories for each condition. The way in which responses are analysed depends upon the judgement (detection, etc.) and the information sought (detection threshold, ranks or central tendency of conditions, psychological "distances" among conditions). Many methods of analysis are available.

6.1.4.2 Numerical categorical judgement methods

A single-stimulus procedure using an 11-grade numerical categorical scale (SSNCS) was studied and compared to graphic and ratio scales. This study, described in Report ITU-R BT.1082, indicates a clear preference in terms of sensitivity and stability for the SSNCS method when no reference is available.

6.1.4.3 Non-categorical judgement methods

In non-categorical judgements, observers assign a value to each image or image sequence shown. There are two forms of the method.

In continuous scaling, a variant of the categorical method, the assessor assigns each image or image sequence to a point on a line drawn between two semantic labels (e.g. the ends of a categorical scale as in Table 3). The scale may include additional labels at intermediate points for reference. The distance from an end of the scale is taken as the index for each condition.

In numerical scaling, the assessor assigns each image or image sequence a number that reflects its judged level on a specified dimension (e.g. image sharpness). The range of the numbers used may be restricted (e.g. 0-100) or not. Sometimes, the number assigned describes the judged level in "absolute" terms (without direct reference to the level of any other image or image sequence as in some forms of magnitude estimation. In other cases, the number describes the judged level relative to that of a previously seen "standard" (e.g. magnitude estimation, fractionation, and ratio estimation).

Both forms result in a distribution of numbers for each condition. The method of analysis used depends upon the type of judgement and the information required (e.g. ranks, central tendency, psychological "distances").

6.1.4.4 Performance methods

Some aspects of normal viewing can be expressed in terms of the performance of externally directed tasks (finding targeted information, reading text, identifying objects, etc.). Then, a performance measure, such as the accuracy or speed with which such tasks are performed, may be used as an index of the image or image sequence.

Performance methods result in distributions of accuracy or speed scores for each condition. Analysis concentrates upon establishing relations among conditions in the central tendency (and dispersion) of scores and often uses analysis of variance or a similar technique.

6.2 Stimulus-comparison methods

In stimulus-comparison methods, two images or sequences of images are displayed and the viewer provides an index of the relation between the two presentations.

6.2.1 General arrangement

The way viewing conditions, source signals, range of conditions and anchoring, the observers, the introduction to the assessment and the presentation of the results are defined or selected is in accordance with § 2.

6.2.2 The selection of test material

The images or image sequences used are generated in the same fashion as in single-stimulus methods. The resulting images or image sequences are then combined to form the pairs that are used in the assessment trials.

6.2.3 Test session

The assessment trial will use either one monitor or two well-matched monitors and generally proceeds as in single-stimulus cases. If one monitor is used, a trial will involve an additional stimulus field identical in duration to the first. In this case, it is good practice to ensure that, across trials, both members of a pair occur equally often in first and second positions. If two monitors are used, the stimulus fields are shown simultaneously.

Stimulus-comparison methods assess the relations among conditions more fully when judgements compare all possible pairs of conditions. However, if this requires too large a number of observations, it may be possible to divide observations among assessors or to use a sample of all possible pairs.